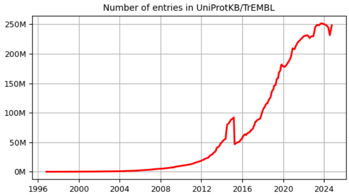
CS612 - Algorithms in Bioinformatics

Databases and Protein Structure Representation

March 3, 2025

Molecular Biology as Information Science

- $> 38,000$ genomes fully sequenced, $> 484,000$ permanent draft, mostly bacterial (2025)
- $254,254,987$ sequences (Nov. 2024), $572,619$ reviewed.
- What do we do with them?
  - Compare them to find what is common and different among organisms (Comparative Genomics)
  - Find out how and which genes encode for which proteins
  - Identify changes that lead to disease
  - Associate structural and functional information with new gene sequences

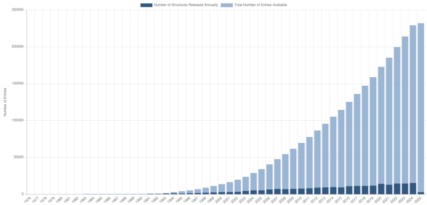


source: `http://www.uniprot.org`

- JGI (Genomes Online Database) `https://gold.jgi.doe.gov/`
- Most of the sequences do not have a solved structure
- Experiments lagging behind
- Way too much data for computer scientists to sit around doing nothing
- Recently – AlphaFold and Large Language Models filling the gap

Molecular Biology as Information Science

- $> 38,000$ genomes fully sequenced, $> 484,000$ permanent draft, mostly bacterial (2025)
- $254,254,987$ sequences (Nov. 2024), $572,619$ reviewed.
- What do we do with them?
  - Compare them to find what is common and different among organisms (Comparative Genomics)
  - Find out how and which genes encode for which proteins
  - Identify changes that lead to disease
  - Associate structural and functional information with new gene sequences



source: `https://www.rcsb.org/stats/growth/growth-released-structures`

- JGI (Genomes Online Database) `https://gold.jgi.doe.gov/`
- Most of the sequences do not have a solved structure
- Experiments lagging behind
- Way too much data for computer scientists to sit around doing nothing
- Recently – AlphaFold and Large Language Models filling the gap

What We Expect From a Biological Databases

- Sequence, functional, structural information, related bibliography
- Well Structured and Indexed
- Well cross-referenced (with other databases)
- Periodically updated and maintained
- Provides tools for analysis and visualization
- Or at least formatted in a compatible way with known tools

Sequence Databases

- International Nucleotide Sequence Database Collaboration (INSDC): `http://www.insdc.org/`
  - NCBI (National Center for Biotechnology Information): `http://ncbi.nih.gov`
  - EMBL-EBI (European Molecular Biology Laboratory, European Bioinformatics Institute): `https://www.ebi.ac.uk/`
  - DDBJ (DNA Data Bank of Japan): `http://www.ddbj.nig.ac.jp/`

Contents of a Database

- Sequences/structures/pathways etc. (depends on the database)
- Accession number
- References
- Taxonomic data
- Annotation/curation
- Keywords
- Cross-reference to relevant data in this or other databases.
- Documentation

Organization of a Database

- Hierarchical, where the data is organized at multiple levels.
- Examples: SCOP, CATH, the tree of life.
- Relational: An entry is a set of correspondences between different features of the database (tables).
- It makes it easy to answer queries using operations like union, intersection, difference etc.

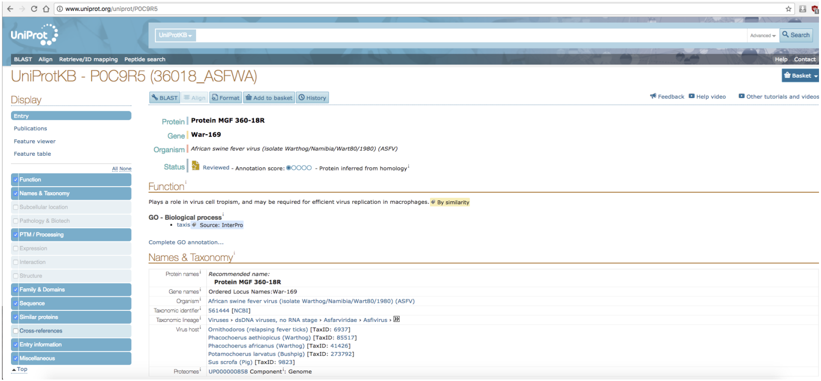
NCBI Nucleotide Sequence Databases

- NCBI GenBank (The nucleotide sequence database) – `http://www.ncbi.nlm.nih.gov/genbank/`
- Provides tools for submission (BankIt, Sequin), retrieval (Entrez) and analysis (BLAST, Genome workbench)
- Provides easy access to other NCBI resources

Protein Sequence Databases

- Uniprot – http://www.uniprot.org/
- A universal resource, resulting from a merger of several databases.
- Tools: BLAST, align, Retrieve/IDmapping
- Pfam – https://www.ebi.ac.uk/interpro/
- A database of protein families based on conserved regions.
- Original site decommissioned in January 2023.
- Now hosted by InterPro.

Uniprot Entry



UniProtKB - P0C9R5 (36018_ASFWA)

Protein	**Protein MGF 360-18R**
Gene	**War-169**
Organism	*African swine fever virus (isolate Warthog/Namibia/Wart80/1980) (ASFV)*
Status	Reviewed - Annotation score: ●○○○○ - Protein inferred from homology

Function

Plays a role in virus cell tropism, and may be required for efficient virus replication in macrophages. By similarity

GO - Biological process

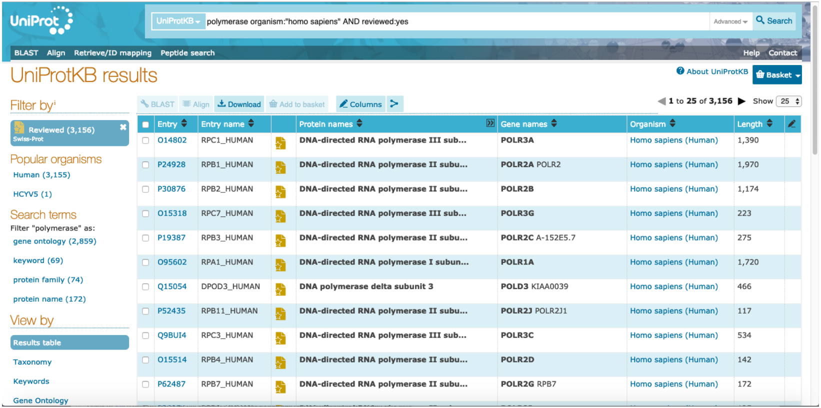
- taxis Source: InterPro

Complete GO annotation...

Names & Taxonomy

Protein names	*Recommended name:* **Protein MGF 360-18R**
Gene names	Ordered Locus Names:War-169
Organism	African swine fever virus (isolate Warthog/Namibia/Wart80/1980) (ASFV)
Taxonomic identifier	561444 [NCBI]
Taxonomic lineage	Viruses › dsDNA viruses, no RNA stage › Asfarviridae › Asfivirus ›
Virus host	Ornithodoros (relapsing fever ticks) [TaxID: 6937] Phacochoerus aethiopicus (Warthog) [TaxID: 85517] Phacochoerus africanus (Warthog) [TaxID: 41426] Potamochoerus larvatus (Bushpig) [TaxID: 273792] Sus scrofa (Pig) [TaxID: 9823]
Proteomes	UP000000858 Component: Genome

Uniprot Search



UniProtKB ▾ polymerase organism:"homo sapiens" AND reviewed:yes   Advanced ▾   Search

BLAST   Align   Retrieve/ID mapping   Peptide search   Help   Contact

UniProtKB results

About UniProtKB   Basket ▾

Filter by

Reviewed (3,156) Swiss-Prot

Popular organisms

- Human (3,155)
- HCYV5 (1)

Search terms

Filter "polymerase" as:

- gene ontology (2,859)
- keyword (69)
- protein family (74)
- protein name (172)

View by

- Results table
- Taxonomy
- Keywords
- Gene Ontology

BLAST   Align   Download   Add to basket   Columns

◀ 1 to 25 of 3,156 ▶   Show 25

□	Entry	Entry name		Protein names	Gene names	Organism	Length
□	O14802	RPC1_HUMAN		DNA-directed RNA polymerase III sub...	POLR3A	Homo sapiens (Human)	1,390
□	P24928	RPB1_HUMAN		DNA-directed RNA polymerase II subu...	POLR2A POLR2	Homo sapiens (Human)	1,970
□	P30876	RPB2_HUMAN		DNA-directed RNA polymerase II subu...	POLR2B	Homo sapiens (Human)	1,174
□	O15318	RPC7_HUMAN		DNA-directed RNA polymerase III sub...	POLR3G	Homo sapiens (Human)	223
□	P19387	RPB3_HUMAN		DNA-directed RNA polymerase II subu...	POLR2C A-152E5.7	Homo sapiens (Human)	275
□	O95602	RPA1_HUMAN		DNA-directed RNA polymerase I subun...	POLR1A	Homo sapiens (Human)	1,720
□	Q15054	DPOD3_HUMAN		DNA polymerase delta subunit 3	POLD3 KIAA0039	Homo sapiens (Human)	466
□	P52435	RPB11_HUMAN		DNA-directed RNA polymerase II subu...	POLR2J POLR2J1	Homo sapiens (Human)	117
□	Q9BUI4	RPC3_HUMAN		DNA-directed RNA polymerase III sub...	POLR3C	Homo sapiens (Human)	534
□	O15514	RPB4_HUMAN		DNA-directed RNA polymerase II subu...	POLR2D	Homo sapiens (Human)	142
□	P62487	RPB7_HUMAN		DNA-directed RNA polymerase II subu...	POLR2G RPB7	Homo sapiens (Human)	172

Protein Structure Databases

- PDB – Protein Data Bank – http://www.rcsb.org/pdb/
- SCOP2 – Structural Classification of Proteins v.2 – http://scop2.mrc-lmb.cam.ac.uk/
- CATH – Another structural classification database – http://www.cathdb.info/
- EMDB – Electron microscopy Database – https://www.ebi.ac.uk/pdbe/emdb/ (Actually part of the PDB now)

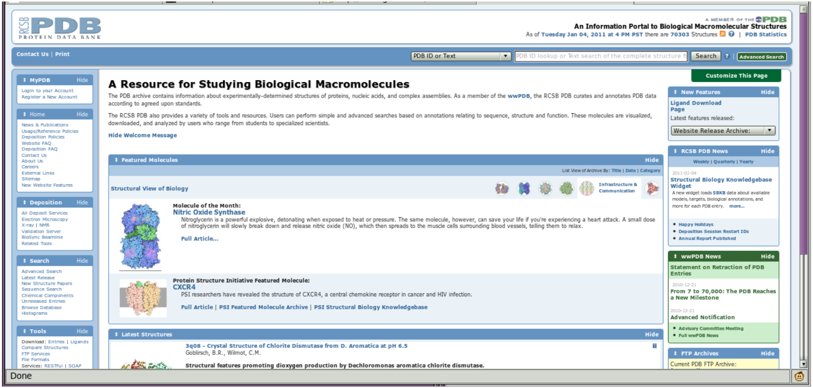
The Protein Databank (PDB)

- Most (all) of the protein structures discovered to date can be found in a large protein repository called the The RCSB Protein DataBank (PDB): `http://www.rcsb.org`.
- PDB is a public domain repository that contains experimentally determined structures of three-dimensional proteins.
- The majority of the proteins in the PDB have been determined by x-ray crystallography.
- The number of proteins determined using NMR methods has been increasing as efficient computational techniques to derive structures from NMR data have been developed.

Retrieving Protein Structures from the PDB

- Starting with 7 structures in 1971, the number has been growing exponentially since then.
- There are approximately 239,000 experimental structures + over a million models as of today (early 2025).
- All PDB entries are 4-letter words! 1CRZ, 2BHL . . .
- Sometimes the chain number is added: 1CRZA, 1CRZB . . .
- You can download the coordinates and display the structure
- The BLAST server and other databases contain links to PDB entries if the sequence has a known structure.

The PDB



An Information Portal to Biological Macromolecular Structures

As of Tuesday Jan 04, 2011 at 4 PM PST there are 70303 Structures | PDB Statistics

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Hide Welcome Message

Featured Molecules Hide

List View of Archive By: Title | Date | Category

Structural View of Biology

Infrastructure & Communication

Molecule of the Month:
Nitric Oxide Synthase

Nitroglycerin is a powerful explosive, detonating when exposed to heat or pressure. The same molecule, however, can save your life if you're experiencing a heart attack. A small dose of nitroglycerin will slowly break down and release nitric oxide (NO), which then spreads to the muscle cells surrounding blood vessels, telling them to relax.

Full Article...

Protein Structure Initiative Featured Molecule:
CXCR4

PSI researchers have revealed the structure of CXCR4, a central chemokine receptor in cancer and HIV infection.

Full Article | PSI Featured Molecule Archive | PSI Structural Biology Knowledgebase

Latest Structures Hide

3q08 - Crystal Structure of Chlorite Dismutase from D. Aromatica at pH 6.5
Goblirsch, B.R., Wilmot, C.M.

Structural features promoting dioxygen production by Dechloromonas aromatica chlorite dismutase.

New Features Hide

Ligand Download Page

Latest features released:

Website Release Archive:

RCSB PDB News Hide

Weekly | Quarterly | Yearly

2011-01-04
Structural Biology Knowledgebase Widget
A new widget loads SBKB data about available models, targets, biological annotations, and more for each PDB entry. more...

- Happy Holidays
- Deposition Session Restart IDs
- Annual Report Published

wwPDB News Hide

Statement on Retraction of PDB Entries

2010-12-21
From 7 to 70,000: The PDB Reaches a New Milestone

2010-12-21
Advanced Notification

- Advisory Committee Meeting
- Full wwPDB News

FTP Archives Hide

Current PDB FTP Archive:

The wwPDB

- In recent years, the major database for macromolecular structures is the worldwide PDB (wwPDB) at `http://www.wwpdb.org/`.
- It is a joint effort of the RCSB, the Protein Data Bank Europe (at the European Bioinformatics Institute, EBI), the Protein Databank Japan (based at Osaka University), and the Biological Magnetic Resonance Data Bank (BMRB).

The PDB Header

```
HEADER    CHROMOSOMAL PROTEIN                     02-JAN-87   1UBQ
TITLE     STRUCTURE OF UBIQUITIN REFINED AT 1.8 ANGSTROMS RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: UBIQUITIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606
KEYWDS    CHROMOSOMAL PROTEIN
EXPDTA    X-RAY DIFFRACTION
AUTHOR    S.VIJAY-KUMAR,C.E.BUGG,W.J.COOK
REVDAT   5   09-MAR-11 1UBQ    1       REMARK
REVDAT   4   24-FEB-09 1UBQ    1       VERSN
REVDAT   3   01-APR-03 1UBQ    1       JRNL
REVDAT   2   16-JUL-87 1UBQ    1       JRNL   REMARK
REVDAT   1   16-APR-87 1UBQ    0
JRNL        AUTH   S.VIJAY-KUMAR,C.E.BUGG,W.J.COOK
JRNL        TITL   STRUCTURE OF UBIQUITIN REFINED AT 1.8 A RESOLUTION.
JRNL        REF    J.MOL.BIOL.                   V. 194   531 1987
JRNL        REFN                   ISSN 0022-2836
JRNL        PMID   3041007
JRNL        DOI    10.1016/0022-2836(87)90679-6
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   S.VIJAY-KUMAR,C.E.BUGG,K.D.WILKINSON,R.D.VIERSTRA,
REMARK   1  AUTH 2 P.M.HATFIELD,W.J.COOK
REMARK   1  TITL   COMPARISON OF THE THREE-DIMENSIONAL STRUCTURES OF HUMAN,
REMARK   1  TITL 2 YEAST, AND OAT UBIQUITIN
REMARK   1  REF    J.BIOL.CHEM.                  V. 262  6396 1987
REMARK   1  REFN                   ISSN 0021-9258
REMARK   1 REFERENCE 2
REMARK   1  AUTH   S.VIJAY-KUMAR,C.E.BUGG,K.D.WILKINSON,W.J.COOK
REMARK   1  TITL   THREE-DIMENSIONAL STRUCTURE OF UBIQUITIN AT 2.8 ANGSTROMS
REMARK   1  TITL 2 RESOLUTION
REMARK   1  REF    PROC.NATL.ACAD.SCI.USA        V.  82  3582 1985
REMARK   1  REFN                   ISSN 0027-8424
REMARK   1 REFERENCE 3
REMARK   1  AUTH   W.J.COOK,F.L.SUDDATH,C.E.BUGG,G.GOLDSTEIN
REMARK   1  TITL   CRYSTALLIZATION AND PRELIMINARY X-RAY INVESTIGATION OF
REMARK   1  TITL 2 UBIQUITIN, A NON-HISTONE CHROMOSOMAL PROTEIN
REMARK   1  REF    J.MOL.BIOL.                   V. 130   353 1979
REMARK   1  REFN                   ISSN 0022-2836
REMARK   1 REFERENCE 4
REMARK   1  AUTH   D.H.SCHLESINGER,G.GOLDSTEIN
REMARK   1  TITL   MOLECULAR CONSERVATION OF 74 AMINO ACID SEQUENCE OF
REMARK   1  TITL 2 UBIQUITIN BETWEEN CATTLE AND MAN
REMARK   1  REF    NATURE                        V. 255   423 1975
REMARK   1  REFN                   ISSN 0028-0836
REMARK   2
REMARK   2 RESOLUTION.    1.80 ANGSTROMS.
```

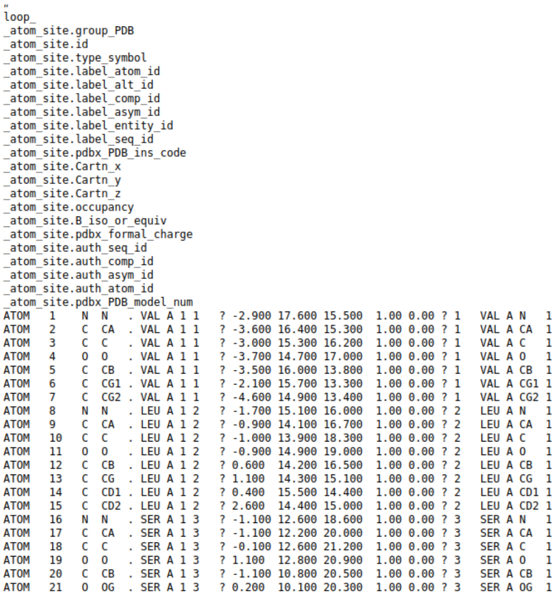
The PDB File Format

```
                          Chain name
    Amino Acid           /   Sequence Number
              \         /   /
     Element   \       /   /     -----Coordinates-----
            \   \     /   /      X       Y       Z      (etc.)
ATOM      1  N   ASP L   1       4.060   7.307   5.186  ...
ATOM      2  CA  ASP L   1       4.042   7.776   6.553  ...
ATOM      3  C   ASP L   1       2.668   8.426   6.644  ...
ATOM      4  O   ASP L   1       1.987   8.438   5.606  ...
ATOM      5  CB  ASP L   1       5.090   8.827   6.797  ...
ATOM      6  CG  ASP L   1       6.338   8.761   5.929  ...
ATOM      7  OD1 ASP L   1       6.576   9.758   5.241  ...
ATOM      8  OD2 ASP L   1       7.065   7.759   5.948  ...
               \\
                Element position within amino acid
```

The PDBx/mmCIF Format

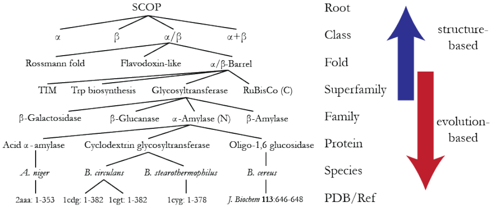
- Developed by the International Union of Crystallography (IUCr) and the Protein Data Bank
- mmCIF is a flexible and extensible tag-value format (dictionary like)
- A newer format designed to address limitations of the PDB format in terms of capacity and flexibility, especially with large structures.
- It is now the default format, and the old format is becoming outdated.
- `https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data`

The PDBx/mmCIF Coordinates



```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N   . VAL A 1 1   ? -2.900 17.600 15.500  1.00 0.00 ? 1   VAL A N   1
ATOM   2    C  CA  . VAL A 1 1   ? -3.600 16.400 15.300  1.00 0.00 ? 1   VAL A CA  1
ATOM   3    C  C   . VAL A 1 1   ? -3.000 15.300 16.200  1.00 0.00 ? 1   VAL A C   1
ATOM   4    O  O   . VAL A 1 1   ? -3.700 14.700 17.000  1.00 0.00 ? 1   VAL A O   1
ATOM   5    C  CB  . VAL A 1 1   ? -3.500 16.000 13.800  1.00 0.00 ? 1   VAL A CB  1
ATOM   6    C  CG1 . VAL A 1 1   ? -2.100 15.700 13.300  1.00 0.00 ? 1   VAL A CG1 1
ATOM   7    C  CG2 . VAL A 1 1   ? -4.600 14.900 13.400  1.00 0.00 ? 1   VAL A CG2 1
ATOM   8    N  N   . LEU A 1 2   ? -1.700 15.100 16.000  1.00 0.00 ? 2   LEU A N   1
ATOM   9    C  CA  . LEU A 1 2   ? -0.900 14.100 16.700  1.00 0.00 ? 2   LEU A CA  1
ATOM   10   C  C   . LEU A 1 2   ? -1.000 13.900 18.300  1.00 0.00 ? 2   LEU A C   1
ATOM   11   O  O   . LEU A 1 2   ? -0.900 14.900 19.000  1.00 0.00 ? 2   LEU A O   1
ATOM   12   C  CB  . LEU A 1 2   ? 0.600  14.200 16.500  1.00 0.00 ? 2   LEU A CB  1
ATOM   13   C  CG  . LEU A 1 2   ? 1.100  14.300 15.100  1.00 0.00 ? 2   LEU A CG  1
ATOM   14   C  CD1 . LEU A 1 2   ? 0.400  15.500 14.400  1.00 0.00 ? 2   LEU A CD1 1
ATOM   15   C  CD2 . LEU A 1 2   ? 2.600  14.400 15.000  1.00 0.00 ? 2   LEU A CD2 1
ATOM   16   N  N   . SER A 1 3   ? -1.100 12.600 18.600  1.00 0.00 ? 3   SER A N   1
ATOM   17   C  CA  . SER A 1 3   ? -1.100 12.200 20.000  1.00 0.00 ? 3   SER A CA  1
ATOM   18   C  C   . SER A 1 3   ? -0.100 12.600 21.200  1.00 0.00 ? 3   SER A C   1
ATOM   19   O  O   . SER A 1 3   ? 1.100  12.800 20.900  1.00 0.00 ? 3   SER A O   1
ATOM   20   C  CB  . SER A 1 3   ? -1.100 10.800 20.500  1.00 0.00 ? 3   SER A CB  1
ATOM   21   O  OG  . SER A 1 3   ? 0.200  10.100 20.300  1.00 0.00 ? 3   SER A OG  1
```

Classification of Protein Structures - The SCOP Database

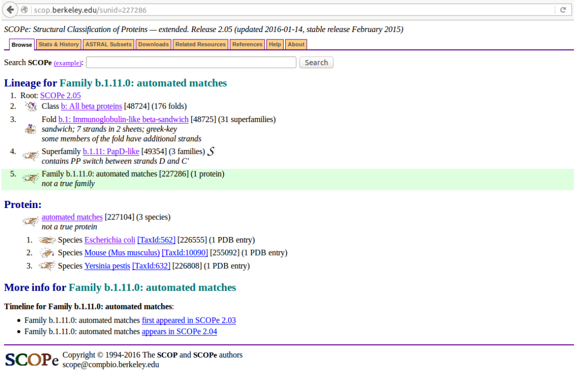


Chothia, Murzin (Cambridge)

Hand-curated hierarchical taxonomy of proteins based on their structural and evolutionary relationships.

- Classes
- Fold Level
- Super Family
- Family
- Domain

The SCOPe Database



SCOPe: Structural Classification of Proteins — extended. Release 2.05 (updated 2016-01-14, stable release February 2015)

Search **SCOPe** (example):

Lineage for Family b.1.11.0: automated matches

1. Root: SCOPe 2.05
2. Class b: All beta proteins [48724] (176 folds)
3. Fold b.1: Immunoglobulin-like beta-sandwich [48725] (31 superfamilies)
   sandwich; 7 strands in 2 sheets; greek-key
   some members of the fold have additional strands
4. Superfamily b.1.11: PapD-like [49354] (3 families) *S*
   contains PP switch between strands D and C'
5. Family b.1.11.0: automated matches [227286] (1 protein)
   not a true family

Protein:

automated matches [227104] (3 species)
not a true protein

1. Species Escherichia coli [TaxId:562] [226555] (1 PDB entry)
2. Species Mouse (Mus musculus) [TaxId:10090] [255092] (1 PDB entry)
3. Species Yersinia pestis [TaxId:632] [226808] (1 PDB entry)

More info for Family b.1.11.0: automated matches

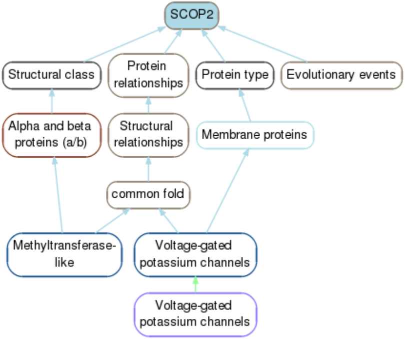
Timeline for Family b.1.11.0: automated matches:

- Family b.1.11.0: automated matches first appeared in SCOPe 2.03
- Family b.1.11.0: automated matches appears in SCOPe 2.04

SCOPe Copyright © 1994-2016 The **SCOP** and **SCOPe** authors
scope@compbio.berkeley.edu

- The successor of SCOP (which is no longer maintained/updated).
- Rather similar, combination of hand-curated and automated methods.

The SCOP2 Database Prototype



- Similar to SCOP(e), but different.
- Adding evolutionary events and protein types among others.
- Several new hierarchical categories.
- The evolutionary relationships induce a graph-like structure rather than rigid hierarchy.

The CATH Database

- Another database which classifies protein structures downloaded from the Protein Data Bank.
- It is a semi-automatic, hierarchical classification of protein domains initially published in 1997.
- CATH is an acronym of the four main levels in the classification.

#	Level	Description
1	**C**lass	Overall secondary-structure content of the domain. (Equivalent to SCOP class)
2	**A**rchitecture	High structural similarity but no evidence of homology. (Equivalent to SCOP fold)
3	**T**opology	A large-scale grouping of topologies which share particular structural features
4	**H**omologous superfamily	Indicative of a demonstrable evolutionary relationship. (Equivalent to SCOP superfamily)

The CATH Database

- Much of the work is done by automatic methods, however there are important manual elements to the classification.
- First – separate the proteins into domains. It is difficult to produce an unequivocal definition of a domain and this is one area in which CATH and SCOP differ.
- The domains are automatically sorted into classes and clustered on the basis of sequence similarities.
- These groups form the **H** levels of the classification. The topology level is formed by structural comparisons of the homologous groups.
- Finally, the **A**rchitecture level is assigned manually.

The CATH Database

Class Level classification is done on the basis of 4 criteria:

1. Secondary structure content;
2. Secondary structure contacts;
3. Secondary structure alternation score; and
4. Percentage of parallel strands.

CATH defines four classes: mostly-α, mostly-β, α and β, few secondary structures.